

# On the Relation Between the Randomized Extended Kaczmarz Algorithm and Coordinate Descent

Bogdan Dumitrescu<sup>a,b</sup>

September 2, 2014

## Abstract

In this note we compare the randomized extended Kaczmarz (EK) algorithm and randomized coordinate descent (CD) for solving the full-rank overdetermined linear least-squares problem and prove that CD needs less operations for satisfying the same residual-related termination criteria. For the general least-squares problems, we show that running first CD to compute the residual and then standard Kaczmarz on the resulting consistent system is more efficient than EK.

**Keywords:** randomized algorithms, least-squares, Kaczmarz method, coordinate descent

**MSC:** 65F10, 65F20, 15A06

<sup>a</sup>Department of Automatic Control and Computers, University Politehnica of Bucharest, Spl. Independenței 313, Bucharest 060042, Romania. E-mail: bogdan.dumitrescu@acse.pub.ro

<sup>b</sup>Department of Signal Processing, Tampere University of Technology, Finland. E-mail: bogdan.dumitrescu@tut.fi

# 1 Introduction

Given matrix  $\mathbf{A} \in \mathbb{R}^{m \times n}$  and vector  $\mathbf{b} \in \mathbb{R}^m$ , the linear least-squares (LS) problem consists of finding  $\mathbf{x} \in \mathbb{R}^n$  such that  $\|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2$  is minimum. Unless explicitly stated, we consider the full-rank overdetermined problem, i.e.  $m \geq n$  and  $\text{rank} \mathbf{A} = n$ . Besides standard solutions based on orthogonal triangularization or the normal equations, significant recent interest was focused on randomized algorithms, showing clear benefits for certain categories of problems, especially for large dimensions and sparse matrices.

There are two main classes of randomized algorithms for the LS problem, both based on simple iterated projection operations. In coordinate descent (CD) [2], at iteration  $k$ , the current residual is projected onto a random column of the matrix  $\mathbf{A}$ , in order to obtain the optimal LS update of a single element of the current approximation of the solution  $\mathbf{x}^{(k)}$ . In the Kaczmarz algorithm [6], the solution  $\mathbf{x}^{(k)}$  is projected onto the hyperplane defined by a random row of the matrix  $\mathbf{A}$  and the respective element of  $\mathbf{b}$ , thus obtaining the next approximation  $\mathbf{x}^{(k+1)}$ . Unlike CD, randomized Kaczmarz converges only when the system  $\mathbf{A}\mathbf{x} = \mathbf{b}$  is consistent. Otherwise, it hovers around the LS solution, within guaranteed bounds [3]. This behavior is natural, since at each iteration the approximated solution satisfies exactly an equation of the system  $\mathbf{A}\mathbf{x} = \mathbf{b}$ , which is not the case in general for the LS solution. Convergence to the LS solution can be achieved if sub-optimal steps are used, see e.g. [1] and the references therein, and the step length goes to zero; however, convergence speed may become very slow.

To fix this drawback, the extended Kaczmarz (EK) algorithm, in both randomized [7] and original deterministic [5] forms, simultaneously builds an approximation of the residual, such that a consistent system is asymptotically obtained, and applies Kaczmarz iterations for the current approximation of this system. Thus, the algorithm converges to the LS solution.

We show in this note that EK consists in fact of CD and Kaczmarz iterations, thus combining both classes of randomized algorithms. Since CD can find on its own the LS solution, we argue that EK can never be faster than CD, neither in terms of number of iterations, nor in terms of number of operations. So, we conclude that randomized CD should be preferred over EK in all situations, for overdetermined LS problems. We discuss also the general LS problem (not full rank) and show that EK can be safely replaced by CD followed by the usual Kaczmarz for better practical behavior. Other combinations of the algorithms are possible for providing early estimates of the solution, like EK.

The notation resembles that from [7]. We denote by  $\mathbf{A}^{(i)}$  and  $\mathbf{A}_{(j)}$  the  $i$ -th row and  $j$ -th column of matrix  $\mathbf{A}$ , respectively, both seen as column vectors. The scalar product of two vectors is denoted  $\langle \cdot, \cdot \rangle$  and  $[m] = \{1, \dots, m\}$ . The  $i$ -th unit vector is  $\mathbf{e}_i$ . To distinguish between algorithms, we add the subscript EK, CD or K (the latter for standard Kaczmarz) to variables having the same meaning, but different values for the three algorithms. We denote  $\mathbf{x}_o$  the solution of the LS problem and  $\mathcal{R}(\mathbf{A})$  the range of  $\mathbf{A}$ . The 2-norm is used for vectors and matrices unless otherwise stated.

## 2 Extended Kaczmarz vs coordinate descent

Algorithm 1 shows a slightly modified version of the randomized EK from [7]. Besides non-significant permutations of the steps and some different explanations, only step 6 is new here and does not affect the final outcome. Let us first discuss the structure of the

---

**Algorithm 1:** Randomized Extended Kaczmarz

---

**Data:** Matrix  $\mathbf{A} \in \mathbb{R}^{m \times n}$ , vector  $\mathbf{b} \in \mathbb{R}^m$ , stopping tolerances  $\varepsilon_{CD}$ ,  $\varepsilon_K$

**Result:** Least-squares solution  $\mathbf{x} \in \mathbb{R}^n$  minimizing  $\|\mathbf{b} - \mathbf{Ax}\|$

- 1 Initialize  $\mathbf{r}_{CD}^{(0)} = \mathbf{b}$ ,  $\mathbf{x}_{CD}^{(0)} = \mathbf{0}$ ,  $\mathbf{x}_{EK}^{(0)} = \mathbf{0}$ .
- 2 **for**  $k = 0, 1, 2, \dots$  **do**
- 3     Pick  $j_k \in [n]$  with probability  $p_j = \|\mathbf{A}_{(j)}\|_2^2 / \|\mathbf{A}\|_F^2$ ,  $j \in [n]$
- 4     Find optimal coordinate descent step:  $\mu = \frac{\langle \mathbf{r}_{CD}^{(k)}, \mathbf{A}_{(j_k)} \rangle}{\|\mathbf{A}_{(j_k)}\|_2^2}$
- 5     Update CD residual:  $\mathbf{r}_{CD}^{(k+1)} = \mathbf{r}_{CD}^{(k)} - \mu \mathbf{A}_{(j_k)}$
- 6     Update CD solution:  $\mathbf{x}_{CD}^{(k+1)} = \mathbf{x}_{CD}^{(k)} + \mu \mathbf{e}_{j_k}$
- 7     Pick  $i_k \in [m]$  with probability  $q_i = \|\mathbf{A}^{(i)}\|_2^2 / \|\mathbf{A}\|_F^2$ ,  $i \in [m]$
- 8     Update EK solution:  $\mathbf{x}_{EK}^{(k+1)} = \mathbf{x}_{EK}^{(k)} + \frac{\langle \mathbf{b} - \mathbf{r}_{CD}^{(k+1)}, \mathbf{e}_{i_k} \rangle - \langle \mathbf{x}_{EK}^{(k)}, \mathbf{A}^{(i_k)} \rangle}{\|\mathbf{A}^{(i_k)}\|_2^2} \mathbf{A}^{(i_k)}$
- 9     Check every  $8 \min(m, n)$  iterations and terminate if both following conditions hold

$$\frac{\|\mathbf{A}^T \mathbf{r}_{CD}^{(k)}\|_2}{\|\mathbf{A}\|_F^2 \|\mathbf{x}^{(k)}\|_2} \leq \varepsilon_{CD} \quad (1)$$

$$\frac{\|\mathbf{b} - \mathbf{r}_{CD}^{(k)} - \mathbf{Ax}^{(k)}\|_2}{\|\mathbf{A}\|_F \|\mathbf{x}^{(k)}\|_2} \leq \varepsilon_K \quad (2)$$


---

algorithm and explain its relation with CD. For further reference, we denote

$$\mathbf{r}^{(k)} = \mathbf{b} - \mathbf{Ax}^{(k)} \quad (3)$$

the residual at iteration  $k$ .

The EK algorithm, as presented in [7], has two intertwined parts. In the first, a residual is built, converging to the optimal residual  $\mathbf{b} - \mathbf{Ax}_o$  of the LS problem. At each iteration, a column  $j_k$  is picked randomly as in step 3 of Algorithm 1, and the residual is projected onto the orthogonal complement of this column, thus obtaining a new residual (smaller in size than the previous because of the projection). However, this is exactly what CD does and that is why we denote this residual  $\mathbf{r}_{CD}^{(k)}$ . Indeed, in CD, the residual is projected onto column  $j_k$  in order to find the optimal update of the  $j_k$ -th element of the solution, as in step 4 (this projection maximizes the decrease of  $\|\mathbf{b} - \mathbf{Ax}^{(k+1)}\|$  if only the  $j_k$ -th coordinate of  $\mathbf{x}^{(k)}$  is modified). After updating the solution as in step 6, the new residual from step 5 is indeed

$$\mathbf{r}_{CD}^{(k+1)} = \mathbf{b} - \mathbf{Ax}_{CD}^{(k+1)} = \mathbf{b} - \mathbf{Ax}_{CD}^{(k)} - \mu \mathbf{A}_{(j_k)} = \mathbf{r}_{CD}^{(k)} - \mu \mathbf{A}_{(j_k)}$$

and is orthogonal on column  $j_k$ , as one can easily check by plugging in the expression of the optimal update  $\mu$ :

$$\langle \mathbf{r}_{CD}^{(k+1)}, \mathbf{A}_{(j_k)} \rangle = \langle \mathbf{r}_{CD}^{(k)} - \mu \mathbf{A}_{(j_k)}, \mathbf{A}_{(j_k)} \rangle = 0.$$

So, steps 3-5 of EK compute the CD residual. Only one more arithmetic operation per iteration is necessary to update the CD solution, as in step 6.

We conclude that Algorithm 1 without steps 7 and 8 is actually the randomized CD algorithm, which converges to the LS solution, i.e.  $\mathbf{x}_{CD}^{(k)} \rightarrow \mathbf{x}_o$ ,  $\mathbf{r}_{CD}^{(k)} \rightarrow \mathbf{b} - \mathbf{A}\mathbf{x}_o$ , see [2], or [4] for a more general treatment. (The probabilities  $p_j$  are also taken like in the randomized CD.) In step 9, CD needs only the stopping criterion (1), which shows that the residual has become nearly orthogonal on  $\mathcal{R}(\mathbf{A})$ . Note that the stopping criterion (2) is irrelevant for CD, since  $\mathbf{b} - \mathbf{r}_{CD}^{(k)} - \mathbf{A}\mathbf{x}_{CD}^{(k)} = \mathbf{0}$  by definition (3). In what follows, we understand by CD the algorithm described in this paragraph, with  $\mathbf{x}_{CD}^{(k)}$  used in (1).

The second part of Algorithm 1, steps 7 and 8, implements a Kaczmarz iteration for the LS problem

$$\mathbf{A}\mathbf{x} = \mathbf{b} - \mathbf{r}_{CD}^{(k)}. \quad (4)$$

Since CD converges to the LS solution, the above system becomes asymptotically consistent and hence the Kaczmarz iterations converge also to the true solution, as shown in [7]. Both stopping criteria from step 9 are now necessary; as above, the criterion (1) shows that the residual has converged; the criterion (2) shows that the Kaczmarz iterations have converged and hence a solution to (4) has been found. In [7], the tolerances  $\varepsilon_{CD}$  and  $\varepsilon_K$  are equal, but we take them different for the sake of generality. Formally, EK is Algorithm 1 with  $\mathbf{x}_{EK}^{(k)}$  used in the stopping criteria (1) and (2) of step 9.

**Remark 1** The above presentation of the CD and EK algorithms allows a quick assessment. CD computes an approximation of the optimal residual of the LS problem and produces an approximation of the LS solution which always satisfies (4). EK computes an approximation of the LS solution by approximating the solution of (4), a system depending on the CD residual. So, EK builds an approximation based on the CD approximation; EK aims to a target that is driven by CD. By its very principle, EK should need more iterations than CD to meet the same stopping criterion. We give below a formal proof of this fact. ■

**Proposition 1** In average (over the random draw of columns and rows), CD terminates in less iterations than EK. Also, CD needs less arithmetic operations than EK.

*Proof.* It is enough to prove the proposition for a literal implementation of Algorithm 1, where the random columns  $j_k$  are the same for EK and CD. Then, by averaging, the same relations hold.

We can safely assume that the EK and CD solution approximations have similar magnitudes, i.e. the values  $\|\mathbf{x}_{EK}^{(k)}\|$  and  $\|\mathbf{x}_{CD}^{(k)}\|$  do not make the stopping criterion (1) significantly different for EK and CD, especially near convergence, the LS solution being unique. The stopping criterion (1) depends only on the CD residual, so EK cannot stop earlier than CD. As mentioned above, the stopping criterion (2) is always met for CD because (4) holds exactly for  $\mathbf{x}_{CD}^{(k)}$ . So, again, EK cannot stop earlier than CD. Hence, CD needs at most the same number of iterations as EK to terminate.

The number of arithmetic operations per iteration is also in favor of CD, since CD needs only a subset of the operations of EK (step 6 is negligible with respect to the others). CD needs about  $4m$  operations per iteration (steps 4 and 5 dictate the complexity), while EK needs about  $4m + 4n$  (steps 4, 5 and 8). ■

Of course, when the CD and EK algorithms are separately implemented, then the random columns are different and it may happen that EK terminates faster than CD, due to a more favorable draw of columns.

Proposition 1 describes the relation between CD and EK from a strictly computational viewpoint, that of algorithm termination. However, the relation between their residuals can be more precisely qualified.

**Proposition 2** Asymptotically, the residuals of CD and EK satisfy  $\|\mathbf{r}_{EK}^{(k)}\| \geq \|\mathbf{r}_{CD}^{(k)}\|$ .

*Proof.* Define

$$\hat{\mathbf{r}}_{EK}^{(k)} = \mathbf{b} - \mathbf{r}_{CD}^{(k)} - \mathbf{A}\mathbf{x}_{EK}^{(k)} \stackrel{(3)}{=} \mathbf{A}(\mathbf{x}_{CD}^{(k)} - \mathbf{x}_{EK}^{(k)}) \quad (5)$$

the residual of the system (4) that EK actually attempts to solve at iteration  $k$ . It results that

$$\mathbf{r}_{EK}^{(k)} = \mathbf{r}_{CD}^{(k)} + \hat{\mathbf{r}}_{EK}^{(k)}. \quad (6)$$

Since CD converges to the solution of the LS problem, its residual tends to become orthogonal to the range of  $\mathbf{A}$ , thus (5) implies that

$$\langle \mathbf{r}_{CD}^{(k)}, \hat{\mathbf{r}}_{EK}^{(k)} \rangle \rightarrow 0.$$

Hence one can infer from (6) that, asymptotically,  $\|\mathbf{r}_{EK}^{(k)}\| \geq \|\mathbf{r}_{CD}^{(k)}\|$ . ■

The above Propositions show that CD reaches its target faster than EK. Of course, a smaller residual does not necessarily mean that the solution approximation is closer to the LS optimum, although this is more likely. In this context, one may wonder if the convergence speed is indeed different for the two algorithms.

**Remark 2** In [7, Th.2.3], the CD residual is shown to satisfy the relation

$$E\{\|\mathbf{r}_{CD}^{(k)} - \mathbf{r}_o\|^2\} \leq \left(1 - \frac{1}{\kappa_F^2(\mathbf{A})}\right)^k \|\mathbf{b} - \mathbf{r}_o\|^2, \quad (7)$$

where  $\mathbf{r}_o = \mathbf{b} - \mathbf{A}\mathbf{x}_o$  is the optimal residual and  $\kappa_F(\mathbf{A}) = \|\mathbf{A}\|_F \|\mathbf{A}^\dagger\|$ , with  $\mathbf{A}^\dagger$  the Moore-Penrose pseudoinverse of  $\mathbf{A}$ . The average in (7) is taken over the random column indices generated in step 3 of Algorithm 1. Since

$$\mathbf{A}(\mathbf{x}_{CD}^{(k)} - \mathbf{x}_o) = \mathbf{r}_o - \mathbf{r}_{CD}^{(k)},$$

it results from (7) that

$$E\{\|\mathbf{x}_{CD}^{(k)} - \mathbf{x}_o\|^2\} \leq \left(1 - \frac{1}{\kappa_F^2(\mathbf{A})}\right)^k \|\mathbf{A}^\dagger\|^2 \cdot \|\mathbf{b} - \mathbf{r}_o\|^2, \quad (8)$$

On the other side, [7, Th.4.1] shows that the EK solution satisfies

$$E\{\|\mathbf{x}_{EK}^{(k)} - \mathbf{x}_o\|^2\} \leq \left(1 - \frac{1}{\kappa_F^2(\mathbf{A})}\right)^{\lfloor k/2 \rfloor} C, \quad (9)$$

where  $C$  is a constant of no interest here.

Although the bounds (8) and (9) are not necessarily tight, they suggest that CD converges faster than EK, supporting the results of Propositions 1 and 2. ■

### 3 The general LS problem

Reminding that all the previous discussion was for full-rank overdetermined LS problems, let us look at the other cases. Consider first underdetermined systems, but still full-rank. In this case, the system  $\mathbf{Ax} = \mathbf{b}$  is consistent and it is well known that the standard Kaczmarz algorithm converges to the LS solution. There is no need of residual approximation, since the residual is zero, hence the CD part of EK is useless.

We pass now to the general LS problem, in which the matrix  $\mathbf{A}$  has arbitrary rank, and for which the deterministic EK algorithm [5] was originally intended. The LS solution is that with minimum norm  $\|\mathbf{x}\|_2$ , among those minimizing the residual  $\|\mathbf{b} - \mathbf{Ax}\|_2$ . In this case, due to their specific projection operations, CD can minimize the residual, but not find a solution with minimum norm, while Kaczmarz can minimize the norm of the solution (if properly initialized with  $\mathbf{x}^{(0)} \in \mathcal{R}(\mathbf{A}^T)$ ), but not that of the residual. EK combines their strengths to find the LS solution.

We argue that, however, there are better ways to combine the two algorithms than intertwining them as in EK. We propose to run first CD for estimating the (nearly) optimal residual  $\mathbf{r}_{CD} \approx \mathbf{r}_o$  and only then Kaczmarz for finding the least norm solution of the consistent system

$$\mathbf{Ax} = \mathbf{b} - \mathbf{r}_{CD}. \quad (10)$$

We name CD+K this algorithm. We note that the general idea of running CD before Kaczmarz is mentioned in [7] (where CD is named "orthogonal projection"); however, the authors settle there for the specific form of EK and discuss CD and K only separately.

**Remark 3** In average, CD+K should need less Kaczmarz iterations than EK. We cannot give a rigorous proof, but give below two heuristic arguments supporting the above assertion.

*Argument 1.* Since the CD operations are independent of the other operations in EK, it takes the same number of iterations for CD and EK to satisfy the stopping criterion (1). Running Kaczmarz after CD has the advantage that it works from the start on the (nearly) consistent system to be solved. In EK, the Kaczmarz iterations are made for the system (4), which is only an approximation of the final consistent system (10).

More intuitively, while CD goes straightly to its target, the Kaczmarz part of EK takes a detour. Running first CD should be more efficient because CD sets the final target, then K goes directly to it. Both CD and K use their full power. So, it is natural to expect less Kaczmarz iterations in CD+K than in EK.

*Argument 2.* Let us take a look at the convergence speed. In [7, Th.3.4], the Kaczmarz algorithm is shown to satisfy the relation

$$E\{\|\mathbf{x}_K^{(k)} - \mathbf{x}_o\|^2\} \leq \left(1 - \frac{1}{\kappa_F^2(\mathbf{A})}\right)^k \|\mathbf{x}_K^{(0)} - \mathbf{x}_o\|^2. \quad (11)$$

The constant bounding the convergence speed is the same as for CD, see (8). So, the discussion from Remark 2 applies also here, suggesting that CD+K has better convergence speed than EK. ■

**Remark 4** One may argue that EK still has an advantage over CD+K: rough approximations of the solution are earlier available. Indeed, in CD+K we have to wait for the whole

CD part before approximations of the solution are computed. A possible fix is to recognize that between EK and CD+K there is a whole family of algorithms, organized as follows.

First CD is run with a tolerance  $\hat{\epsilon}_{CD} > \epsilon_{CD}$ . Then EK is run, initialized with the residual produced by CD, until one of the stopping criteria (1) or (2) is satisfied. If (2) is satisfied we stop, otherwise we run Kaczmarz on the now nearly consistent system, initializing with the solution produced by EK, until (2) is met. We name CD+EK+K this algorithm. Again, we expect it to be faster than EK, the arguments being similar to those in Remark 3. If  $\hat{\epsilon}_{CD}$  is large, only few CD iterations are made, hence approximations of the solution are quickly available. ■

**Remark 5** Running Kaczmarz after CD has a nice alternative interpretation. In this context, the Kaczmarz iteration has the form (see step 8 of Algorithm 1)

$$\mathbf{x}_K^{(k+1)} = \mathbf{x}_K^{(k)} + \frac{\langle \mathbf{b} - \mathbf{r}_{CD}, \mathbf{e}_{i_k} \rangle - \langle \mathbf{x}_K^{(k)}, \mathbf{A}^{(i_k)} \rangle}{\|\mathbf{A}^{(i_k)}\|_2^2} \mathbf{A}^{(i_k)} = \mathbf{x}_K^{(k)} + \frac{\langle \mathbf{x}_{CD} - \mathbf{x}_K^{(k)}, \mathbf{A}^{(i_k)} \rangle}{\|\mathbf{A}^{(i_k)}\|_2^2} \mathbf{A}^{(i_k)}, \quad (12)$$

where we have used the equality  $\langle \mathbf{b} - \mathbf{r}_{CD}, \mathbf{e}_{i_k} \rangle = \langle \mathbf{A}\mathbf{x}_{CD}, \mathbf{e}_{i_k} \rangle = \langle \mathbf{x}_{CD}, \mathbf{A}^{(i_k)} \rangle$ . Denoting  $\mathbf{q}_K^{(k)} = \mathbf{x}_{CD} - \mathbf{x}_K^{(k)}$ , it results from (12) that

$$\mathbf{q}_K^{(k+1)} = \mathbf{q}_K^{(k)} - \frac{\langle \mathbf{q}_K^{(k)}, \mathbf{A}^{(i_k)} \rangle}{\|\mathbf{A}^{(i_k)}\|_2^2} \mathbf{A}^{(i_k)}. \quad (13)$$

This is a projection operation on the orthogonal complement of the  $i_k$ -th row of  $\mathbf{A}$ , dual to the CD operation of projecting the residual on the orthogonal complement of column  $j_k$  (step 5 of Algorithm 1). Hence,  $\mathbf{q}_K^{(k)}$  tends to the component of  $\mathbf{q}_K^{(0)}$  that is orthogonal on  $\mathcal{R}(\mathbf{A}^T)$ . Initializing with  $\mathbf{q}_K^{(0)} = \mathbf{x}_{CD}$ , which corresponds to the natural initialization  $\mathbf{x}_K^{(0)} = 0$ , the iteration (13) converges to  $\mathbf{q}_K$  satisfying  $\mathbf{x}_K + \mathbf{q}_K = \mathbf{x}_{CD}$ , with  $\mathbf{q}_K \perp \mathcal{R}(\mathbf{A}^T)$  and hence  $\mathbf{x}_K \in \mathcal{R}(\mathbf{A}^T)$ . Since  $\mathbf{A}\mathbf{q}_K = 0$ , the Kaczmarz solution satisfies the system (10). This means that  $\mathbf{x}_K = \mathbf{x}_o$ , since the LS solution is the unique vector from  $\mathcal{R}(\mathbf{A}^T)$  satisfying (10).

So, the iteration (13) starts with  $\mathbf{x}_{CD}$ , for which (10) already holds, and computes its projection onto the orthogonal complement of  $\mathcal{R}(\mathbf{A}^T)$ . Thus, it allows the computation of the projection of  $\mathbf{x}_{CD}$  onto  $\mathcal{R}(\mathbf{A}^T)$ , which is the LS solution.

Of course, using (13) instead of (12) gives no computational advantage, but gives a dual view to the convergence of the Kaczmarz iterations. ■

## 4 Numerical results

We have implemented Algorithm 1 in Matlab and report the performance of CD and EK only in terms of number of iterations, reminding however that at similar number of iterations CD is still faster. The algorithm has been run for a sufficiently high number of iterations, without any stopping criterion. For overdetermined LS problems, we have considered matrices belonging to two classes where EK was shown in [7] to have better performance than other algorithms: (i) dense well-conditioned matrices, generated with `randn`, and (ii) sparse random matrices with density 0.25, generated with `sprandn`.

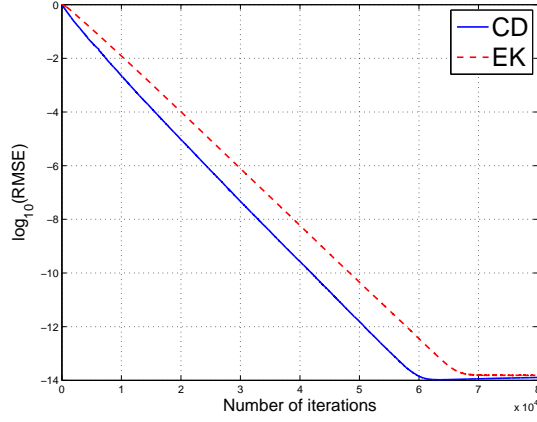


Figure 1: RMSE for dense matrices,  $m = 2000$ ,  $n = 500$ .

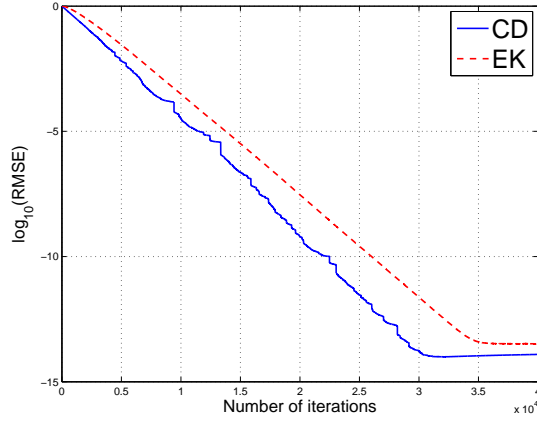


Figure 2: RMSE for dense matrices,  $m = 10000$ ,  $n = 500$ .

We report the normalized RMSE  $\sqrt{E(\|x^{(k)} - x_o\|^2 / \|x_o\|^2)}$ , obtained by averaging over 100 matrices from the same class. Figures 1 and 2 show the RMSE for dense matrices with the same number of rows,  $n = 500$ , but different number of columns:  $m = 2000$  and  $m = 10000$ , respectively. Figure 3 shows the RMSE for sparse matrices,  $n = 800$ ,  $m = 2000$ . In all cases, the faster convergence of CD is clear. When the system is very overdetermined, CD has a jagged convergence, alternating many small advances with few large ones, but is still faster. For other matrix sizes, the results are similar.

To illustrate the behavior of CD+K and CD+EK+K, we have generated random matrices  $A \in \mathbb{R}^{m \times n}$ , computed their SVD  $A = U\Sigma V^T$ , kept only the  $r$  largest singular values in  $\Sigma$  while setting the others to zero, and recomputed  $A$  using the same above relation. We have implemented CD+EK+K in a simple manner that ensures a simple evaluation: first CD is run for  $N$  iterations, then EK for a number of iterations (that may vary depending on matrix size and rank), then K is run for  $N$  iterations. Note that CD and K run the same number of iterations, like in EK.

We give only one typical sample of result, for an underdetermined problem with  $m = 500$ ,  $n = 2000$ ,  $r = 400$ . Figure 4 shows the RMSE of the solution approximations as a



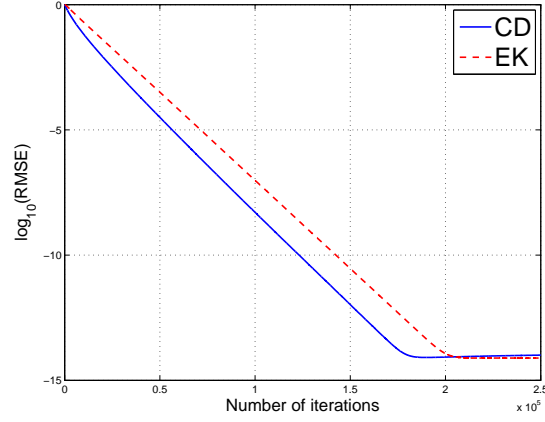


Figure 3: RMSE for sparse matrices,  $m = 2000$ ,  $n = 800$ .

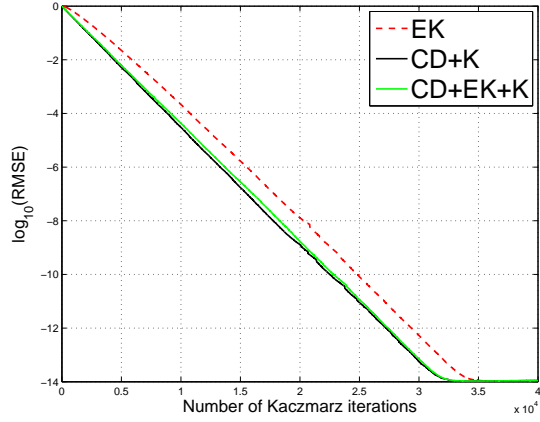


Figure 4: RMSE for dense matrices,  $m = 500$ ,  $n = 2000$ ,  $N = 2000$ .

function of the number of Kaczmarz iterations; the number of CD iteration is the same for all methods, but their position in time is different. It is visible that both CD+K and CD+EK+K need less Kaczmarz iterations to converge. Although  $N = 2000$  for CD+EK+K, which is a relatively small value, the performance is about the same as for CD+K; for larger values, like  $N = 5000$ , the curves for CD+K and CD+EK+K are nearly identical; for smaller values, the curve of CD+EK+K approaches that of EK.

All the presented curves show that the advantage of CD, CD+K or CD+EK+K over EK is built especially in the first iterations. This corresponds well with the fact that in those iterations the approximation of the residual is still very poor in EK, and hence the Kaczmarz iterations do not have a good target, as explained in Remark 1 and in argument 1 of Remark 3.

## 5 Conclusions

The computational conclusion of all the facts presented in this note is the recommendation to replace EK with one of the following algorithms:

- CD, for full-rank overdetermined LS problems;
- CD+K or CD+EK+K, for rank-deficient or unknown rank problems.

(As already known, Kaczmarz replaces EK for full-rank underdetermined problems.)

In particular, for the full-rank overdetermined problem, CD is always preferable to EK, due to the following reasons:

- In average, CD converges in less iterations than EK.
- CD needs less operations per iteration.
- CD uses only the columns of the matrix  $\mathbf{A}$ , while EK uses both columns and rows.

The conclusions apply equally to randomized EK [7] and the less efficient deterministic version [5].

## Acknowledgment

The author is indebted to Liang Dai for stimulating discussions on the Kaczmarz algorithm and for pointing out [7].

## References

- [1] M. Hanke and W. Niethammer. On the Acceleration of Kaczmarz's Method for Inconsistent Linear Systems. *Lin. Alg. Appl.*, 130:83–98, 1990.
- [2] D. Leventhal and A. S. Lewis. Randomized Methods for Linear Constraints: Convergence Rates and Conditioning. *Math. Oper. Res.*, 35(3):641–654, Aug. 2010.
- [3] D. Needell. Randomized Kaczmarz solver for noisy linear systems. *BIT Numer. Math.*, 50:395–403, 2010.
- [4] Yu. Nesterov. Efficiency of coordinate descent methods on huge scale optimization problems. *SIAM J. Optim.*, 22(2):341–362, 2012.
- [5] C. Popa. Least-squares solution of overdetermined inconsistent linear systems using Kaczmarz's relaxation. *Int. J. Comput. Math.*, 55:79–89, 1995.
- [6] T. Strohmer and R. Vershynin. A Randomized Kaczmarz Algorithm with Exponential Convergence. *J. Fourier Anal. Appl.*, 15:262–278, 2009.
- [7] A. Zouzias and N.M. Freris. Randomized Extended Kaczmarz for Solving Least Squares. *SIAM J. Matrix Anal. Appl.*, 34(2):773–793, 2013.